Text-Free Prosody-Aware Generative Spoken Language Modeling

Year: 2021 | Citations: 136 | Authors: E. Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet

Abstract

Speech pre-training has primarily demonstrated efficacy on classification tasks, while its capability of generating novel speech, similar to how GPT-2 can generate coherent paragraphs, has barely been explored. Generative Spoken Language Modeling (GSLM) (CITATION) is the only prior work addressing the generative aspect of speech pre-training, which builds a text-free language model using discovered units. Unfortunately, because the units used in GSLM discard most prosodic information, GSLM fails to leverage prosody for better comprehension and does not generate expressive speech. In this work, we present a prosody-aware generative spoken language model (pGSLM). It is composed of a multi-stream transformer language model (MS-TLM) of speech, represented as discovered unit and prosodic feature streams, and an adapted HiFi-GAN model converting MS-TLM outputs to waveforms. Experimental results show that the pGSLM can utilize prosody to improve both prosody and content modeling, and also generate natural, meaningful, and coherent speech given a spoken prompt. Audio samples can be found at https://speechbot.github.io/pgslm. Codes and models are available at https://github.com/pytorch/fairseq/tree/main/examples/textless_nlp/pgslm.